Automatic selection of Unobserved Components models for supply chain forecasting.

Marco A. Villegas, Diego J. Pedregal

ETSI Industriales de Ciudad Real. Universidad de Castilla-La Mancha.

Abstract

For many companies, automatic forecasting has come to be an essential part of Business Analytics applications. The large amounts of data available, the short life-cycle of the analysis and the acceleration of business operations make traditional handmade data analysis unfeasible in such environments. In this paper, an Automatic Forecasting Support System comprising several methods and models is developed in a general State Space framework built in the so called SSpace toolbox written for Matlab. Some of the models included such as Exponential Smoothing and ARIMA are well-known, but we propose a new model family that has very rarely been used before in this context, namely Unobserved Components models. Additional novelties are that Unobserved Components models are used in an automatic identification environment and that their forecasting performance is compared with Exponential Smoothing and ARIMA models estimated with different software packages. The daily sold units dataset of a franchise chain in Spain spanning 166 products and 517 days of sales is used to assess empirically the new system. The system works well in practice and the proposed automatic Unobserved Components models compare very favorably with other methods and other well-known software packages in forecasting terms.

Keywords: Automatic forecasting, model selection, State Space, Unobserved Components, Kalman Filter, Business Analytics

1. Introduction

Business Analytics (BA) combines the latest tools and methodologies to apply statistical and computational modeling and leverage organizational data to enhance the decision making process in business administration and management science. Apart from the multiple technical challenges posed in a real BA implementation, one particular difficulty consists of selecting the most appropriate model to analyze a given dataset. This problem is becoming more and more important as BA has to come to grips with requirements such as online and real-time service levels in big data scenarios.

A disconnect has always existed between two centers of attention in this field. At one extreme lies a set of business concepts and abstractions (e.g., revenue, costs and margin), while at the other are the raw data generated by daily business operations (e.g., point-of-sale or manufacturing transactions). A huge conceptual gap still separates these two focal points, although some recent efforts have been made to solve the issue, including some conceptual modeling approaches (Decker et al., 2010; Barone et al., 2010). Most propose *conceptual mapping* between business-concepts and raw data-entities. While all these attempts have been both laudable and essential, when dealing with real BA implementations we also have to address the issue of selecting the right statistical technique to best analyze and model the raw data. This polarization means that the success of a good BA implementation inevitably relies on the identification of appropriate data modeling techniques, which is considered to be an unavoidable intermediate step to bridge the two extremes.

In this context, forecasting models are strategic in nature because they are at the root of business decisions ranging from inventory management and scheduling to strategic management (Petropoulos et al., 2014). Focusing on the supply chain context explored in this paper, automatic model selection is essential due to the high number of products whose demand has to be forecast (Fildes and Petropoulos, 2015). In such an envi-

[1]ETSI Industriales. Universidad de Castilla-La Mancha. Edificio Politécnico. 13071 - Ciudad Real.

ronment, the distinction between aggregate versus individual model selection is very relevant (Fildes, 1989). In the aggregate approach, a single source of forecasts is chosen for all the time series, while in the individual a particular method for each series is selected on the basis of a forecasting criteria. In general terms, individual selection outperforms aggregate selection, but at the cost of higher complexity level and computational burden (Fildes and Petropoulos, 2015).

Automatic model selection has received a great deal of attention in the specialized time series literature. This interest spans from classical modeling techniques, such as regression analysis, Exponential Smoothing, ARIMA, Transfer Functions, etc. (Hocking, 1976; Gómez and Maravall, 2001; Hyndman et al., 2008; Young, 2011), to more modern Big Data techniques, such as Artificial Neural Networks, Support Vector Machines, etc. (Haykin, 2008; Sapankevych and Sankar, 2009; Hastie et al., 2009). It even might be said that, while classical methods traditionally used to rely on crafted detailed identification tools supervised by humans, more modern models inevitably have to rely on automatic identification and selection techniques due to the amount of data that has to be processed.

In fact, the advent of the Big Data era has meant that even traditional techniques have had to adapt to the times, with companies accumulating so many time series to model and forecast that detailed manual identification becomes literally impossible (Fildes and Petropoulos, 2015). In addition, the short life-cycle of the analysis and the acceleration of business operations contribute to the need for automatic identification techniques.

Automatic identification techniques very often rely on a statistical criterion that is systematically applied to a subspace of feasible models, with the selection of the best model depending on the criterion in question. Many of these criteria sprang up in the 1970's, and have been modified at a later date (Peña, 2001). The Hannan and Quinn Criterion, the Akaike Information Criterion and Schwarz's Bayesian Criterion are often reported in the literature as sensible and reliable statistical criteria (Akaike, 1974; Hannan and Quinn, 1979; Schwarz, 1978). This point of view is well grounded when models are considered correct representations of the data, but have been challenged by a more agnostic point of view based on evidence that, in real life data, in-sample optimal performance according to any of the above criteria does not necessarily imply a superior out-of-sample forecasting performance (Fildes and Petropoulos, 2015).

Exponential Smoothing (ETS) continues to be the most used modeling technique in business and industry, at least in areas ranging from inventory management and scheduling to planning and strategic management. Its endurance over time is striking, especially as it started as an ad hoc method for smoothing and forecasting time series in the 1950's (Gardner, 1985, 2006). The main reasons for its longevity are probably its good performance in general applications and its intuitive appeal, as it is easy to understand and to communicate to company managers. The fact that they were not grounded in statistics has fueled a root and branch revision in the last 20 years that has dramatically changed the view held of ETS techniques, including automatic identification tools and software (Hyndman et al., 2008; Hyndman and Khandakar, 2008).

The second most widespread methodology after ETS is, to all appearances, ARIMA. More recent than ETS, ARIMA models have expanded since the 1970's after the publication of the influential book by Box and Jenkins (Box et al., 2015). Different automatic identification procedures have been proposed (Gómez and Maravall, 2001; Hyndman and Khandakar, 2008), with TRAMO (in conjunction with SEATS) probably being the most used ARIMA automatic identification procedure in statistical agencies all over the world.

At the same time, the academic literature on the topic has evolved much more than its practical implementation in business and industry, which for the most part remain staunch advocates of ETS and ARIMA. Numerous papers can be found that add new dimensions to the problem, such as judgmental forecasting, modeling the bullwhip effect, intermittent demands, promotions, etc. (Croston, 1972; Syntetos and Boylan, 2001; Trapero et al., 2013, 2015; Trapero and Pedregal, 2016). However, many other papers focus on comparing other techniques, such as Artificial Neural Networks, models with inputs, etc. Fildes et al. (2008); Syntetos et al. (2009, 2016) provide comprehensive surveys.

Despite the wealth of literature on the topic, there is a family of models that has been systematically overlooked, namely structural Unobserved Components models (UC) (Harvey, 1989; Young et al., 1999; Durbin and Koopman, 2012). UC have not been tested in the present context for at least four reasons. First, UC have been developed in academic environments, with no strategy for their dissemination among practitioners for their everyday use in business and industry. They have been applied to a certain degree in state agencies with modest success in terms of dissemination, due to the overwhelming use of X-12-ARIMA and its predecessors, TRAMO and SEATS. Second, the widely-held but not-scientifically-tested feeling that UC models do not really have anything relevant to add to ETS methods has deterred its use in practice (UC "bear many similarities to exponential smoothing methods, but have multiple sources of random error. In particular, the 'basic structural model' (BSM) is similar to Holt-Winters' method for seasonal data and includes level, trend and seasonal components", see De Gooijer and Hyndman (2006)). Third, UC models are usually identified by hand, with automatic identification being very rare. Finally, software is scarcer and sparser than standard packages, some genuine alternatives are e.g., STAMP (Koopman et al., 1999), SSfpack (Koopman et al., 2008), SSpace (Villegas and Pedregal, 2016).

A good demonstration of the lack of use of UC models compared to other methods is that they are not included in many forecasting competitions and studies of forecasting comparisons in supply chain contexts. The list is too long to quote here, but some examples are Makridakis and Hibon (2000a); Fildes et al. (2008); Syntetos et al. (2009); Athanasopoulos et al. (2011); Syntetos et al. (2016). Some of the procedures in these references do in fact involve a comprehensive search among many possibilities as the application of an expert system, but none includes UC models.

The alleged similarities between UC and ETS deserve a special comment. There is no doubt that there are similarities, but such a broad statement requires a more rigorous evaluation. In short, ETS and UC share common ground in the same way that ETS and ARIMA, or UC and ARIMA, do. The key point is that UC have their own avenues of development, with some areas that clearly overlap with ETS and ARIMA, while other areas are quite distinct, which is what makes each of the techniques useful in its own way.

The main advantages of ETS methods according to Hyndman et al. (2008); Hyndman and Khandakar (2008) are, for example, their intuitive appeal, their ability to handle multiplicative-additive mixed models in a very efficient way, the use of a linear innovation state space model and the development of an automatic identification procedure for different combinations of trend models, seasonals and noises.

Without taking anything away from ETS, UC models are much more flexible in several ways, and it is hard to imagine any ETS modification that cannot be implemented in a UC context (indeed, some improvements in UC may also be back replicable in ETS). For example, mixed multiplicative-additive models could be replicated as a UC model with the aid of a non linear state space model and the Extended Kalman Filter and Smoother, although at the cost of a higher computational burden. A further example is that the single source of noise assumed in ETS is in fact an unnecessary constraint, at least seen from the point of view of UC. Why the noise of trends and seasonal components ought to be the same? The point is especially puzzling, because well-known efficient solutions are available for multiple noise models (Durbin and Koopman, 2012). Nevertheless, the single noise structure is perfectly implementable in a UC model, if such an option is considered absolutely essential.

UC also have further advantages. First, there are no limitations on the type of trends (or 'levels') one may choose for a model. Although most typical trends are Local Level (or Random Walk) or a Local Linear Trend, other types are Integrated Random Walk, Double Integrated Random Walk, Smoothed Trend or Smoothed Random Walk (Harvey, 1989). Second, the same is true of the seasonal component: the simplest is a dummy seasonal, but trigonometric or Dynamic Harmonic Regression (DHR) with common or different hyper-parameters for the harmonics involved can also be implemented. Spectral tools are very useful in this respect (Young et al., 1999). Modulated seasonality may be very interesting in certain applications (Pedregal and Young, 2006). Third, several cycles may be introduced as trigonometric or DHR components. Once again, spectral analysis is very valuable in this case. The mean periods of the cycles may be estimated jointly with other parameters if required (Harvey, 1989). Fourth, extending univariate UC to include inputs is straightforward - a topic that seems very relevant to model outliers and the effect of sales promotions. These can be entered as linear regression terms with constant or time varying parameters, transfer functions, or even non linear functions (Harvey, 1989). Finally, multivariate UC models are also available (Durbin and Koopman, 2012).

In order to contribute to the dissemination of UC models to a wider audience, in this paper a general Automatic Forecasting Support System is built for the 166 final products of a food franchise chain with more than 100 stores throughout Spain. Although developed on this dataset, the system can be applied as is to other datasets, and consists of fitting a set of models built on some traditional models such as ETS and ARIMA models with different levels of complexity. The novelty of the system, however, is that UC are also included as a valuable alternative option. All the models in the system are cast in a common State Space framework that is new in two ways. First, as far as the authors know, this is the first occasion on which UC models have been used in a supply chain forecasting context. Second, it is the first time that automatic identification of UC has been proposed. This paper shows that UC are in fact a very powerful choice for forecasting and that they may indeed be identified automatically, which is a standard feature of many other competitors nowadays.

The remainder of the paper is organized as follows: Section 2 presents the forecasting methods in the system. Section 3 gives a brief summary of the State Space general framework in which all of these are actually cast. Section 4 presents specific aspects of the automatic identification routines used. Section 5 sets out the case study and the findings. Finally, section 6 offers some final remarks and conclusions.

2. Forecasting methods

2.1. Benchmarks

Two methods usually found in the literature are tested here as base benchmarks to be improved on: the seasonal Naïve and Autoregressive models (AR). The seasonal Naïve produces forecasts that are simple projections of the last seasonal span into the future.

AR models seek to capture the correlation structure of a time series by regressing a stationary output variable on its past values. Because of the stationarity condition, regular and seasonal difference operators are usually applied to the initial non-stationary data (Box et al., 2015). A general AR(p) model is:

$$\begin{aligned} y_t &= (1-B)^d(1-B^7)^D z_t \\ y_t &= \sum_{i=1}^{p} \phi_i y_{t-i} + a_t = \frac{1}{(1-\phi_1 B - \ldots - \phi_p B^p)} a_t \end{aligned} \tag{1}$$

where a_t is a white noise sequence (i.e., serially uncorrelated with zero mean and constant variance); B stands for the backshift operator so $B^l z_t = z_{t-l}$; $(1-B)$ is the difference operator; $(1-B^7)$ is the seasonal difference operator (the power of B is 7 in this case, as the data is daily with a weekly seasonal); d and D are the regular and seasonal integer difference orders necessary to induce stationarity; ϕ_i are a set of parameters to estimate; and y_t is the stationary series, obtained from differencing the original, non-stationary data, z_t. A constant should be included in the model if the mean of the differenced series y_t is not zero.

2.2. Autoregressive Integrated Moving Avarage model (ARIMA)

This is an extension of the previous model to include multiplicative seasonality and moving average terms, popularized by Box et al. (2015). The formulation of an ARIMA$(p,d,q)\times(P,D,Q)_7$ model for the stationary time series y_t is:

$$y_t = \frac{(1+\theta_1 B + \ldots + \theta_q B^q)}{(1+\phi_1 B + \ldots + \phi_p B^p)} \frac{(1+\Theta_1 B^7 + \ldots + \Theta_Q B^{7Q})}{(1+\Phi_1 B^7 + \ldots + \Phi_P B^{7P})} a_t \tag{2}$$

where new parameters appear, namely $\theta_i (i = 1,2,\ldots,q)$, $\Theta_j (j = 1,2,\ldots,Q)$ and $\Phi_k (k = 1,2,\ldots,P)$.

2.3. Exponential smoothing (ETS)

The main principle in exponential smoothing is that forecasts are built as a correction of the last data by a weighted average of past forecast errors that decay following an exponential pattern. This pattern is applied to reflect the fact that the most recent information is more valuable for forecasting the immediate future than the oldest data. The simplest version is Brown's single exponential smoothing (Brown, 1959), shown in equation (3), where a hat on a variable stands for a forecast of the variable.

$$\hat{z}_{t+1} = \hat{z}_t + \alpha(z_t - \hat{z}_t) \tag{3}$$

This simple model may be completed in many different ways before reaching the well-known Holt-Winters method with seasonality (Winters, 1960). Exponential smoothing techniques have recently been revised in depth and extended in many respects. Some of these additions include the use of damped trends, mixing multiplicative and additive components in one single model, using a State Space framework, and even considering multivariate versions. A detailed description of most of these alternatives can be found in Hyndman et al. (2008).

2.4. Unobserved components models (UC)

These models decompose any time series into parts with economic meaning, such as trends, cycles, seasonals, irregulars, etc. The general formulation used for the data in this paper is

$$z_t = T_t + S_t + I_t \tag{4}$$

where T_t, S_t and I_t stand for the trend, seasonal and irregular components, respectively. The flexibility of this approach is introduced via the components' dynamic specification.

Trends are usually taken from the following family:

$$\begin{bmatrix} T_{t+1} \\ T^*_{t+1} \end{bmatrix} = \begin{bmatrix} \alpha & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} T_t \\ T^*_t \end{bmatrix} + \begin{bmatrix} \eta_t \\ \eta^*_t \end{bmatrix} \tag{5}$$

where T^*_t is referred to as the slope of the trend, $0 < \alpha \leq 1$, η_t and η^*_t are white noise sequences with variances σ^2_η and $\sigma^2_{\eta^*}$, respectively. This model is called Smoothed Trend (ST) and subsumes the following specific cases: i) Random Walk (RW), by eliminating the second equation and $\alpha = 1$; ii) Integrated Random Walk (IRW) with $\alpha = 1$ and $\sigma^2_\eta = 0$; iii) Smoothed Random

Walk (SRW) with $\sigma_\eta^2 = 0$; and iv) Local Linear Trend (LLT) with $\alpha = 1$, see e.g., Harvey (1989), Taylor et al. (2007) and Durbin and Koopman (2012).

Seasonal components used in this paper take a trigonometric form that is more flexible than the dummy seasonal model (Harvey, 1989). The basic component for the $j-th$ harmonic is the following:

$$\begin{bmatrix} S_{t+1}^{(j)} \\ S_{t+1}^{*(j)} \end{bmatrix} = \begin{bmatrix} cos\lambda_j & sin\lambda_j \\ -sin\lambda_j & cos\lambda_j \end{bmatrix} \begin{bmatrix} S_t^{(j)} \\ S_t^{*(j)} \end{bmatrix} + \begin{bmatrix} \psi_t^{(j)} \\ \psi_t^{*(j)} \end{bmatrix} \quad (6)$$

where $\lambda_j = 2\pi j/s$ is the frequency of the seasonal component, with s being the seasonal period or the number of observations per year, $S_t^{*(j)}$ an additional state necessary for the specification, and $\psi_t^{(j)}$ and $\psi_t^{*(j)}$ white noises with common variance σ_j^2.

The full seasonal component is built by block concatenation of this type of model for the fundamental frequency and all its harmonics, i.e., $\lambda_j, j = 1, 2, \ldots, [s/2]$, where $[s/2] = s/2$ for even s numbers, and $[s/2] = (s-1)/2$ for odd s numbers. For the daily data with a weekly seasonal of the case study below $s = 7$ and $j = 1, 2, 3$. The standard stable specification for the seasonal component selects a common noise variance for the fundamental frequency and all its harmonics, i.e., $\sigma_j^2 = \sigma^2, j = 1, 2, 3$.

The above UC model specification sometimes leaves some serial correlation in the innovations. This problem may be tackled in two complementary ways, namely, by introducing higher flexibility in the seasonal component by allowing different variances for each harmonic (Young et al., 1999), or/and selecting a colored noise for the irregular white noise component (Harvey, 1989).

2.5. Combination of methods

The combination of forecasts has received considerable attention in the literature with overwhelming evidence that combining forecasts improves forecasting performance, see e.g., Makridakis and Hibon (2000b); Kourentzes et al. (2014). Despite many different methods existing for combining forecasting methods, often the simplest alternatives, such as mean and median, are the optimal choice, see Barrow and Kourentzes (2016) and references therein. Mean and median are included in this study for the sake of comparison.

3. State Space methods

All the univariate models in the previous section are suitable for the general linear and Gaussian State Space framework, whose general formulation is given by equation (7).

$$\begin{array}{ll} \text{Transition equation:} & \alpha_{t+1} = \Phi_t \alpha_t + E_t \eta_t \\ \text{Observation equation:} & z_t = H_t \alpha_t + C_t \epsilon_t \end{array} \quad (7)$$

In this equation, α_t is the state vector; η_t and ϵ_t are the state and observation noises, drawn from Gaussian distributions with zero mean and variance Q_t and R_t, respectively. The remaining elements in (7) are referred to as the system matrices with appropriate dimensions.

Given model (7), the well-known Kalman Filter and state smoother produce the optimal estimates (in the sense of minimizing the mean squared errors) of the first- and second-order moments (mean and covariance) of the state vector, conditional on all the data in a sample. Exact initialization of the algorithms is available for both stationary and non stationary versions of the general system (7), implying that the initial conditions issue is efficiently solved for all the individual models.

The unknown parts in the system matrices can be estimated by Exact Maximum Likelihood computed using the KF via "prediction error decomposition", see details in Harvey (1989), Young et al. (1999), Taylor et al. (2007), Durbin and Koopman (2012).

All the forecasting techniques used in this paper are tractable with the SSpace toolbox written in MATLAB (Villegas and Pedregal, 2016), available from the authors upon request. In fact, they are simply a subset of a much wider range of linear, non-Gaussian and nonlinear models implementable in SSpace.

4. Automatic model selection

This section briefly summarizes the specific algorithms for automatic model selection for each of the family models considered in this paper. All of these follow the general approach of searching in a wide model space with a given criterion, Schwarz Bayesian Criterion (SBC) in our case, combined in some cases with some data pre-processing, as specified below. The desired result is a parsimonious model, albeit sufficiently complex to be a correct statistical representation of the data.

4.1. AR and ARIMA

The automatic identification of ARIMA models is carried out in two steps in a modified version of Hyndman and Khandakar (2008). The first step consists of identifying the regular and seasonal integration orders (d and D). In the second step, the model space is searched in an intelligent way using SBC.

The way in which the integration orders of time series is handled in Hyndman and Khandakar (2008) is by means of formal unit root statistical tests. Applied to our dataset, these formal tests repeatedly produced wrong seasonal integration orders, leading to forecasts that were misaligned with the true data. The problem was easily solved by replacing the unit root tests with a heuristic rule for selecting the combination of regular and seasonal differences that minimizes the resulting stationary series' variance. The rest of the algorithm follows the indications of Hyndman and Khandakar (2008).

4.2. ETS

The model space is all possible combinations of level, growth, seasonal and noise of the following options: i) simple exponential, double exponential and damped trend for level and growth; ii) seasonal and non seasonal; iii) AR models of orders 0 (white noise), 1, or 2.

4.3. UC

The model space in this case is composed of all combinations of the following possibilities: i) Trends: RW, IRW, LLT, ST, SRW; ii) non seasonal component, seasonal components with common variance for all harmonics, and seasonal component with different variances for all of these; iii) irregular components, such as AR models of order 0 (white noise), 1, 2.

As far as the authors know, this is the first time that UC models have been used in an automatic identification context of the type found in this paper.

5. Case study

The proposed evaluation of the models was performed on a set of 166 products from a food franchise with more than 100 stores throughout Spain. The company specializes in selling everyday dishes made from natural products at affordable prices in take-away and take-in formats. Demand for all these products is not intermittent, in the sense that all values are continuous and well above zero. 517 daily sales observations were made for each product with 414 observations used for in-sample and the remainder for out-of-sample evaluation. In-sample operations consisted of the automatic identification and estimation of all forecasting methods using initial daily observations, and the production of 1 to 14 days ahead forecasts. Therefore, a set of 90, 14 days ahead forecast rounds was carried out for each product. Figure 1 shows some typical examples of the time series in the dataset.

Several metrics are used to evaluate the forecasting performance of each method, namely the average relative Mean Absolute Error (AvgRelMAE), the Absolute Scaled Error (ASE) and the Cumulative Absolute Scaled Error (CASE), see Davidenko and Fildes (2013) and Davidenko and Fildes (2016).

Calling $r_i = \frac{MAE_i^A}{MAE_i^B}$ the ratio of the MAE of model A and horizon h over the MAE of benchmark model B for time series i, the AvgRelMAE for horizon h and m time series is calculated as

$$\text{AvgRelMAE}_h = \exp\left(\frac{1}{m}\sum_{i=1}^{m} \ln r_i\right) \tag{8}$$

AvgRelMAE shows how the adjustments improve/reduce the MAE compared to the benchmark (Naïve in later examples) statistical forecast. Values below 1 means that on average the model outperforms the benchmark. Values above 1 indicates the opposite.

The Absolute Scaled Error with origin T and forecast horizon h (ASE_h) is defined as

$$\text{ASE}_h = \frac{|z_{T+h} - \hat{z}_{T+h}|}{\frac{1}{T}\sum_{i=1}^{T} z_i} \tag{9}$$

The denominator is included to normalize the errors so as to be able to summarize the results across different products. Results in the case study are robust to replacing the absolute error with squared errors in equation (9) and are available from the authors upon request. Taking ASE_h as the basic metrics, the mean and median can be considered valid measures for aggregated forecast performance across all products.

Finally, the Cumulative Absolute Scaled Error with origin T and forecast horizons from 1 to h (CASE_h) is defined as the cumulative mean of ASE_h up to horizon h, i.e.,

$$\text{CASE}_h = \frac{1}{h}\sum_{l=1}^{h} \frac{|z_{T+l} - \hat{z}_{T+l}|}{\frac{1}{T}\sum_{i=1}^{T} z_i} \tag{10}$$

Figure 2 shows the maximum and minimum CASE_{14} values for all models across all individual products, excluding the Naïve, to avoid distortions. There is considerable variation in the forecastability of each product, since there are some products for which errors are below 0.25, while there are a few other products for which errors are above 0.8. The relevance of Figure 2 is that, although differences in forecast errors are detected across models and products, all models behave similarly for

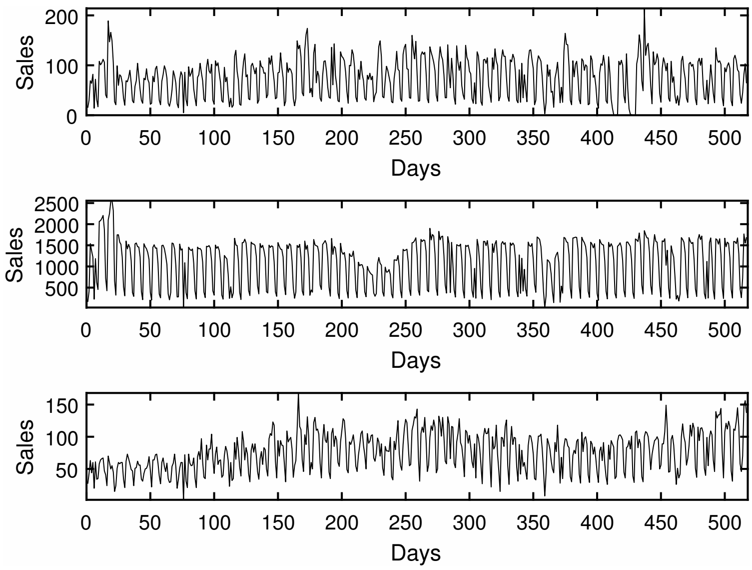

Figure 1: Three examples of demand time series

Table 1: Percentage of products for which each model is best and worst according to $CASE_{14}$.

	Best	Worst
ARIMA	41.0%	5.4%
AR	5.4%	72.3%
ETS	11.4%	17.5%
UC	42.2%	4.9%

the same product in that, when forecast errors are high, all the models show high errors, and vice versa. This ensures that the overall forecasting results are not influenced by extreme values or any other sort of anomaly in the data or in the identification or estimation procedures.

Table 1 complements Figure 2 by showing the percentage of products for which each method is best or worst for all forecasting origins according to $CASE_{14}$. UC and ARIMA compete for top place. ETS takes middle spot, a long way behind, and the AR model is the clear loser.

Tables 2 and 3 show the AvgRelMAE, the median and 75% percentile of ASE_h for all products, all forecast origins and some selected forecast horizons. Figure 3 shows the median and 75% percentile of $CASE_h$ in a graphical form excluding the Naïve model.

Several conclusions may be extracted from this information:

- As expected, forecasting performance worsens with longer horizons.
- All methods outperform both the Naïve and AR benchmarks by a wide margin. Although AR is only considered here as a benchmark, it systematically outperforms Naïve.
- UC is the best model for all forecast horizons, followed by ARIMA and ETS.
- Simple forecast combination methods outperform ARIMA and ETS but do not improve on UC. The performance of ARIMA and ETS is very similar in this case, mainly due to the absence of anomalous data.
- The median of $CASE_h$ and ASE_h (see middle panel of Figure 3 and Table 3) shows that, although UC is a better option, all the models, with the exception of AR and Naïve (not included in the Figure) offer very similar results. This means that it is hard to find differences in the models with performance below the median values. However, as the bottom panel shows, differences are more apparent when the 75% percentile is considered. The advantage of UC models is clearer in these cases.

A robustness test checks for the independence of the results from the software used. Part of the experiment was repeated with other software to run this test, namely the FORECAST package (Hyndman and Khandakar, 2008) in R that allows for automatic identification of AR, ARIMA and ETS models, and TRAMO (Gómez and Maravall, 2001), which allows for an additional ARIMA identification and forecasting system with outlier handling, if desired. The tested in this paper are AR,

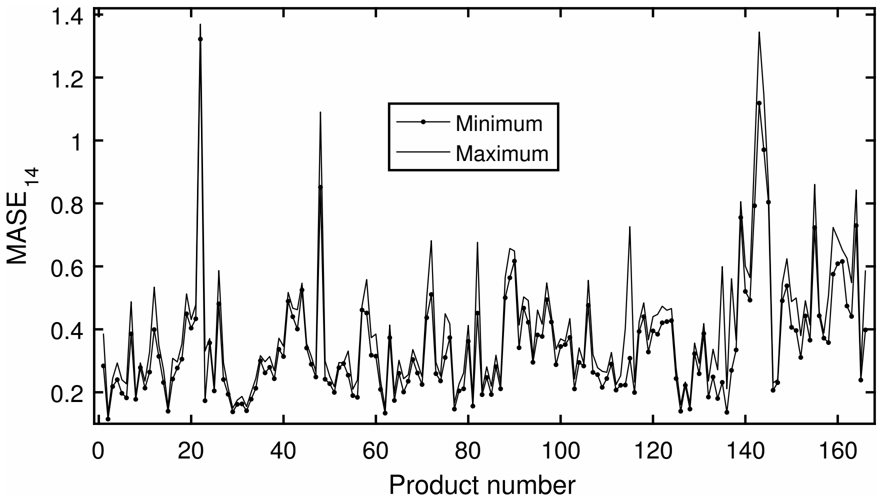

Figure 2: Minimum and maximum values of $CASE_{14}$ for four models and each product.

Table 2: AvgRelMAE for selected forecasting horizons. Boldface used to highlight minimum values of each column.

AvgRelMAE	1	2	3	4	5	6	7	10	14
AR	0.807	0.854	0.869	0.880	0.887	0.890	0.890	0.913	0.915
ARIMA	0.740	0.783	0.801	0.814	0.821	0.824	0.825	0.829	0.839
ETS	0.756	0.802	0.817	0.828	0.833	0.835	0.836	0.844	0.847
UC	**0.730**	**0.772**	**0.790**	**0.801**	**0.808**	**0.811**	**0.813**	**0.817**	**0.824**
Mean	0.735	0.778	0.795	0.807	0.814	0.818	0.820	0.824	0.831
Median	0.733	0.776	0.792	0.804	0.810	0.814	0.815	0.821	0.828

Table 3: Median and 75% percentile of ASE_h for selected forecasting horizons.

Median	1	2	3	4	5	6	7	10	14
NAIVE	0.270	0.270	0.269	0.269	0.270	0.270	0.270	0.279	0.282
AR	0.238	0.252	0.253	0.257	0.259	0.259	0.261	0.272	0.276
ARIMA	0.216	0.228	0.232	0.235	0.237	0.236	0.238	0.243	0.249
ETS	0.219	0.234	0.239	0.241	0.240	0.240	0.241	0.248	0.252
UC	**0.211**	**0.223**	**0.228**	**0.230**	**0.231**	**0.231**	**0.232**	**0.237**	**0.243**
Mean	0.216	0.227	0.231	0.233	0.236	0.237	0.239	0.246	0.248
Median	0.215	0.227	0.231	0.233	0.234	0.234	0.236	0.242	0.246
75% perc.	1	2	3	4	5	6	7	10	14
NAIVE	0.573	0.574	0.572	0.574	0.577	0.577	0.579	0.590	0.592
AR	0.459	0.490	0.503	0.511	0.521	0.522	0.521	0.549	0.551
ARIMA	0.429	0.458	0.470	0.474	0.478	0.482	0.485	0.494	0.504
ETS	0.440	0.471	0.479	0.486	0.485	0.487	0.489	0.506	0.508
UC	**0.422**	**0.443**	**0.457**	**0.460**	**0.463**	**0.467**	**0.469**	**0.481**	**0.483**
Mean	0.426	0.453	0.464	0.468	0.475	0.478	0.481	0.490	0.499
Median	0.428	0.452	0.464	0.470	0.475	0.475	0.479	0.492	0.494

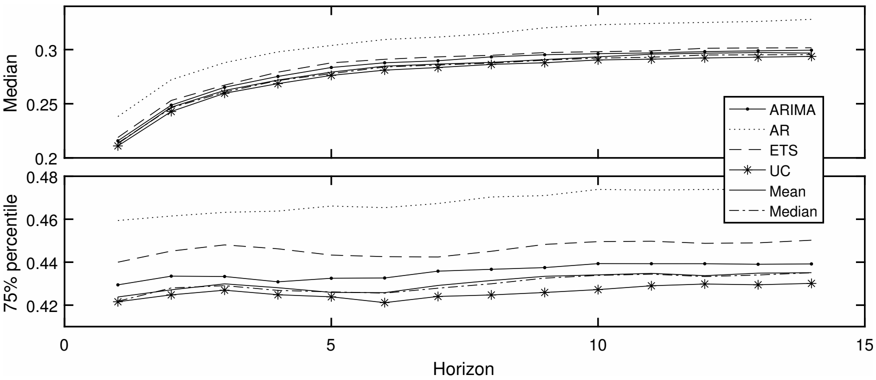

Figure 3: Median (top panel) and 75% percentile (bottom panel) of $CASE_h$ for all models excluding the Naïve.

ARIMA and ETS in R (hereafter with a prefixed 'R-' to distinguish them from the previous models in SSpace) and ARIMA estimated with TRAMO. This last option includes two versions, with and without outlier treatment, referred to as TRAMO(*) and TRAMO, respectively.

Results are summarized in Tables 4, 5 and Figure 4 in similar formats to those used above. Previous UC model performance results are repeated to facilitate comparisons.

Some interesting findings are:

- According to AvgRelMAE, the UC model outperforms all others. Only minor exceptions appear when the median of ASE_h are considered.

- R-AR performs worse than its AR counterpart in the preceding tables and is the second worst method so far, even outperformed by Naïve for horizons greater than 2.

- On the other hand, R-ETS outperforms its ETS equivalent in the preceding tables. This may be due to the fact that R-ETS allows for general ARMA models for the observational noise, while ETS constrain them to AR up to order 2.

- Results shown for R-ARIMA are obtained constraining the seasonal differences to 1 in the automatic procedure. The errors are considerably larger when this parameter is selected fully automatically with formal unit roots. For example, the median of ASE_1 without the constraint is 0.350 compared to 0.217 in Table 5, and the median of ASE_{14} is 0.383 (compared to 0.239 in Table 5).

- TRAMO and TRAMO(*) are worse than R-ARIMA and R-ETS for longer horizons, although better for shorter ones. TRAMO(*) is somewhat better than TRAMO.

- Once more, the median of error measures reduces the magnitude of the errors (see middle panel of Figure 4), i.e., all procedures work similarly for the more easily forecastable cases with the exception of R-AR. Differences among models are more apparent when the 75% percentiles is considered. This is also the case for TRAMO and TRAMO(*), implying that there is a more visible outlier effect in less forecastable cases.

Another important procedure to assess how great differences are between methods is to use formal statistical tests to evaluate whether differences are statistically significant. This can be done with rank tests (see Koning et al. (2005) and references therein) that rank methods by forecast error measurement. Tests are performed in several different ways:

- Overall: the null hypothesis assumes all methods are statistically indistinguishable.

- Multiple comparisons: having rejected the previous hypothesis, this additional test provides a basis for testing which method(s) is(are) responsible for the differences by performing a battery of pair comparison tests. All null hypotheses assume that differences between methods are statistically insignificant. A simple graphical device easily and simultaneously provides evidence of all the tests by presenting the mean rank for each method with their confidence bands. Methods whose intervals do not overlap are significantly different. Forecasting accuracy is the same for methods with overlapping intervals. This graphical evidence for our

Table 4: AvgRelMAE for additional models.

AvgRelMAE	1	2	3	4	5	6	7	10	14
R-AR	0.958	0.991	1.000	1.004	1.004	1.003	1.003	1.007	1.004
R-ARIMA	0.757	0.797	0.810	0.820	0.826	0.829	0.827	0.835	0.840
TRAMO	0.739	0.787	0.807	0.823	0.832	0.837	0.840	0.853	0.865
TRAMO(*)	0.733	0.779	0.796	0.812	0.822	0.828	0.832	0.845	0.861
R-ETS	0.743	0.811	0.816	0.828	0.835	0.834	0.829	0.842	0.833
UC	**0.730**	**0.772**	**0.790**	**0.801**	**0.808**	**0.811**	**0.813**	**0.817**	**0.824**

Table 5: Median and 75% percentile of ASE_h of additional models.

Median	1	2	3	4	5	6	7	10	14
R-AR	0.261	0.269	0.269	0.268	0.270	0.270	0.271	0.283	0.283
R-ARIMA	0.217	0.227	**0.228**	0.232	0.234	0.235	0.234	0.239	0.239
TRAMO	0.214	0.230	0.234	0.238	0.240	0.240	0.240	0.250	0.255
TRAMO(*)	**0.210**	0.225	**0.228**	**0.230**	0.233	0.233	0.233	0.243	0.247
R-ETS	0.218	0.239	0.239	0.243	0.244	0.243	0.238	0.247	0.245
UC	0.211	**0.223**	**0.228**	**0.230**	**0.231**	**0.231**	**0.232**	**0.237**	**0.243**
75% perc.	1	2	3	4	5	6	7	10	14
R-AR	0.543	0.566	0.571	0.580	0.581	0.581	0.580	0.591	0.594
R-ARIMA	0.432	0.455	0.464	0.471	0.476	0.477	0.477	0.492	0.498
TRAMO	0.426	0.455	0.472	0.482	0.484	0.489	0.492	0.508	0.521
TRAMO(*)	0.422	0.450	0.459	0.464	0.471	0.477	0.482	0.501	0.512
R-ETS	0.429	0.469	0.470	0.477	0.484	0.485	0.480	0.499	0.494
UC	**0.422**	**0.443**	**0.457**	**0.460**	**0.463**	**0.467**	**0.469**	**0.481**	**0.483**

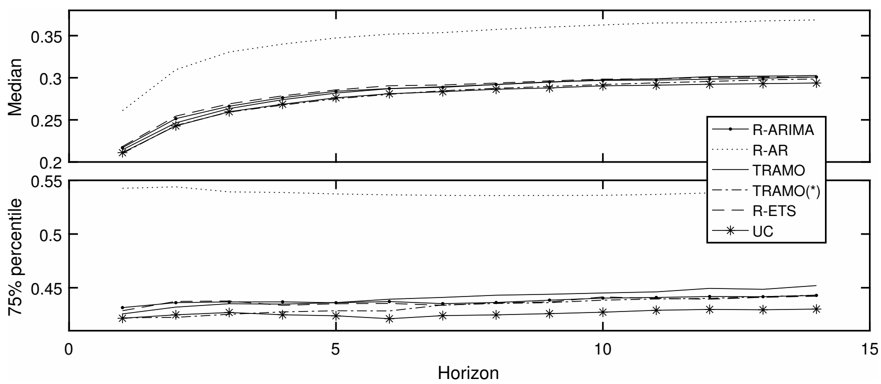

Figure 4: Median (top panel) and 75% percentile (bottom panel) of $CASE_h$ for alternative models.

dataset and a 5% significance level interval is presented in the top panel of Figure 5, where the dotted horizontal line indicates the point at which the best method (namely UC) is worse. In other words, any method with an interval crossed by this line is indistinguishable from the best option, while all the methods with a non-overlapping interval are significantly worse.

- Multiple comparisons with the mean: all methods are compared to mean behavior to determine whether they are better or worse than the mean (based on a null of equal behavior). Once more, a simple graph presents the mean ranks along with the overall mean and its confidence interval. Methods above the confidence band are significantly worse than the mean, while methods below the confidence band are better. Evidence for the data set is presented in the bottom panel of Figure 5.

The evidence provided by Figure 5 is complemented with Table 6. Said table repeats the tests disaggregated by forecast horizon, showing the number of forecasting horizons for which each method is worse or better than the best (first and second column) and the number of horizons for which each method is worse than the mean (third column). Some points emerge from all this evidence:

- The overall test disregarding the benchmark methods (Naïve, AR and R-AR) clearly indicates that methods are significantly different. According to Koning et al. (2005) the statistical value is 106.79 and the critical value for the test at 5% significance level should be sought on a Chi-squared with $K-1$ degrees of freedom, with K being the number of methods (9 in our case). The critical value is 15.51, see details in Koning et al. (2005). The same overall tests done by forecast horizon give values that range between a minimum of 65.77 for a lag of 7, to a maximum of 88.1 for lag 14. Therefore, the null hypothesis of all methods performing equally is rejected in all cases with a high level of confidence.

- Multiple comparisons suggest that UC is significantly different from all other methods except ARIMA and Median. One particular point of interest is that UC is different from both ETS and R-ETS, implying that UC is not as similar to ETS methods as might be expected (see introduction).

- One striking fact in the top panel of Figure 5 is the differences observed between ARIMA, R-ARIMA, TRAMO and TRAMO(*), which are supposed to be methodologically equivalent. ARIMA seems different to TRAMO and TRAMO(*), with R-ARIMA sharing part of its interval on both sides. Differences in automatic identification, initialization issues and optimizer routines are responsible for the difference in performance.

- Multiple comparisons with the mean shown in the bottom panel of Figure 5 reinforce what has already been said. R-ETS and TRAMO perform worse than the mean, while ARIMA and UC perform significantly better.

- Table 6 provides still more evidence along the same lines. The first column shows that, with the exception of ARIMA and Median, most methods are worse than UC most of the time. Similarly, the only methods systematically better and never worse than the mean are ARIMA and UC. However, while the remaining models are never better than the mean, they are sometimes worse.

The evidence provided so far shows clearly that UC is a powerful tool that may be added to the tool box of researchers and practitioners in the area of supply chain forecasting. The reasons why UC is superior in this case is because of its greater flexibility at least in three ways: i) multiple noise State Space systems are used, in particular noises for trends and seasonals are different; ii) an ample menu of trends are trained and iii) seasonal components are more flexible, since models with different variances for the harmonics are allowed. The dataset used in these experiments exhibits strong seasonal components in comparison to the trends, and the advantage of using different sources of noise for different components and the greater flexibility of the seasonal components themselves in UC models is clearly highlighted by their outperformance.

6. Conclusions

The advent of the Big Data era poses many challenges to researchers and practitioners in the field of forecasting in industrial and business environments. One of these challenges is the efficient and robust automatic handling of massive databases. Naturally, this is also essential for the specific area of supply chain forecasting, for which many solutions have been proposed in the academic literature, although everyday applications simply use a much smaller number of techniques.

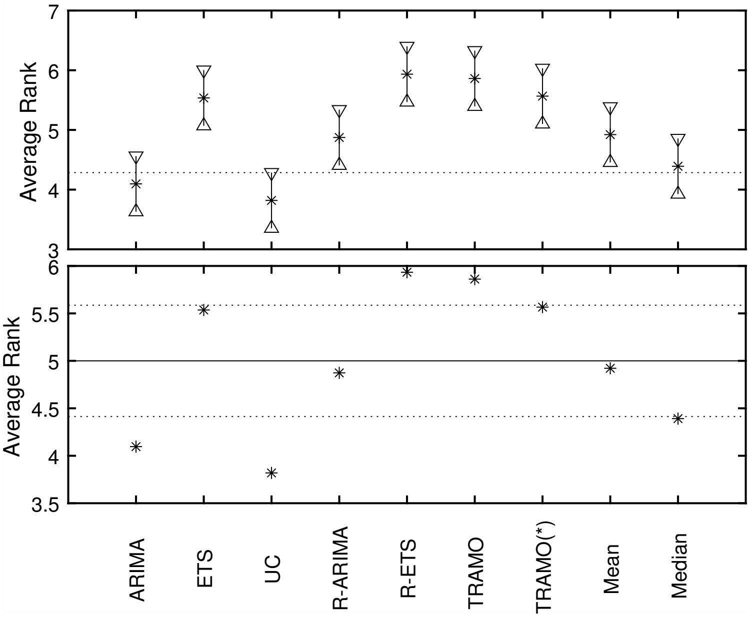

Figure 5: Multiple comparisons plot (top) and multiple comparisons with the mean (bottom) of some forecasting methods.

Table 6: Results of multiple comparisons by forecast horizons.

	No. of horizons with method significantly worse than UC	No. of horizons with method significantly better than UC	No. of horizons with method significantly worse than mean
ARIMA	0	12	0
ETS	14	0	3
UC	0	14	0
R-ARIMA	9	0	1
R-ETS	13	0	8
TRAMO	14	0	9
TRAMO(*)	13	0	4
Mean	12	0	0
Median	0	2	0

In this paper, a general Automatic Forecasting Support System is built in the unified framework provided by State Space methods developed in the SSpace toolbox written in MATLAB. This system takes advantage of well-established automatic identification methods for well-established model families, mainly Exponential Smoothing and ARIMA models. It also incorporates Unobserved Components models, which are quite uncommon in these types of applications.

As far as the authors know, this is the first time UC models have been proposed in a supply chain forecasting context, and this is also the first time automatic identification of UC models has been proposed.

A comprehensive evaluation shows that UC models compete very favorably with other standard alternatives in forecasting terms. This is true for all the models estimated by the toolbox, implying that UC is slightly better than ARIMA and significantly better than ETS.

The results are consistent when the same models are estimated with alternative, well-established software packages, which points to the quality of the software developed and the usefulness of UC models as forecasting devices in the area.

Most importantly, significant differences are also shown to exist among methods. In particular, UC is significantly better than ETS, invalidating, or at least discrediting, the idea that UC models are "similar" to ETS. However, UC is also better than ARIMA models estimated with other software, but very similar to ARIMA in the toolbox.

The findings are clearly specific to the dataset used in this paper and we make no claims whatsoever that they will always be valid for other datasets. Obviously, the approach and the UC models themselves need to be tested on many more datasets for such a conclusion to be drawn. Notwithstanding, this paper provides a starting point for considering UC models as a competitive method with great potential in this and other fields.

One important point is that UC methods' outperformance is not due to any greater complexity in the models themselves. In fact, the State Space framework in which UC models are cast shares common ground with the other methods used. The sole difficulty lies in the fact that they are simply not as well known by academics and practitioners in the field. Therefore, academics and practitioners are cordially invited to experience the benefits of working with UC and/or SSpace for themselves.

7. Acknowledgements

The authors would like to thank Dr. J.R. Trapero for his insightful suggestions on a previous draft of this paper.

This work was supported by the European Regional Development Fund and Spanish Government (MINECO/FEDER, UE) under the project with reference DPI2015-64133-R and by the Vicerrectorado de Investigación y Política Científica from UCLM by DOCM 31/07/2014 [2014/10340].

8. References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, (19):716–723.

Athanasopoulos, G., Hyndman, R., Song, H., and Wu, D. (2011). The tourism forecasting competition. *International Journal of Forecasting*, 27(3):822–844.

Barone, D., Mylopoulos, J., Jiang, L., and Amyot, D. (2010). The Business Intelligence Model: Strategic Modelling.

Barrow, D. K. and Kourentzes, N. (2016). Distributions of Forecasting Errors of Forecast Combinations: Implications for Inventory Management. *International Journal of Production Economics*, (177):24–33.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 5th edition.

Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*. McGraw-Hill.

Croston, J. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 23:289–303.

Davidenko, A. and Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. *International Journal of Forecasting*, 29:510–522.

Davidenko, A. and Fildes, R. (2016). Forecast error measures: Critical review and practical recommendations. In *Business Forecasting: Practical Problems and Solutions*. John Wiley & Sons Inc.

De Gooijer, J. and Hyndman, R. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22:443–473.

Decker, G., Dijkman, R., Dumas, M., and García-Bañuelos, L. (2010). The Business Process Modeling Notation. In *Modern Business Process Automation*, pages 347–368. Springer.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press, 2nd edition.

Fildes, R. (1989). Evaluation of aggregate and individual forecast method selection rules. *Management Science*, 35(9):1056–1065.

Fildes, R., Nikolopoulos, N., Crone, S., and A.A., S. (2008). Forecasting and operational research: a review. *Journal of the Operational Research Society*, 59:1150–1172.

Fildes, R. and Petropoulos, F. (2015). Simple Versus Complex Selection Rules for Forecasting Many Time Series. *Journal of Business Research*, (8):1692–1701.

Gardner, E. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, 4:1–28.

Gardner, E. (2006). Exponential smoothing: The state of the art - part ii. *International Journal of Forecasting*, 22:637–666.

Gómez, V. and Maravall, A. (2001). Automatic Modeling Methods for Univariate Series. In *A Course in Time Series*, pages 171–201. John Wiley & Sons, Inc.

Hannan, E. J. and Quinn, B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, Series B*, (41):190–195.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge university press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer Publishing.

Haykin, S. (2008). *Neural networks and learning machines.* New Jersey: Prentice Hall.

Hocking, R. R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, (32):1–49.

Hyndman, R., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: the State Space Approach*. Springer Science & Business Media.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: The Forecast Package for R. *Journal of Statistical Software*, 3(27):1–22.

Koning, A., Franses, P., Higon, M., and Stekler, H. (2005). The m3 competition: Statistical tests of the results. *International Journal of Forecasting*, 21:397–409.

Koopman, S., Harvey, A., Doornik, J., and Shephard, N. (1999). *Structural Time Series Analysis, Modelling, and Prediction using STAMP.* London: Timberlake Consultants Press.

Koopman, S., Shephard, N., and Doornik, J. (2008). *Statistical Algorithms for Models in State Space Form: SsfPack 3.0*. Timberlake Consultants Press.

Kourentzes, N., Barrow, D., and Crone, S. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, 41(9):4235–4244.

Makridakis, S. and Hibon, M. (2000a). The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16:451–476.

Makridakis, S. and Hibon, M. (2000b). The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476.

Peña, D. (2001). Prediction and Model Selection. In *A Course in Time Series*, pages 111–135. John Wiley & Sons, Inc.

Pedregal, D. and Young, P. (2006). Modulated cycles, an approach to modelling periodic components from rapidly sampled data. *International Journal of Forecasting*, 22:189–194.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., and Nikolopoulos, K. (2014). 'Horses for Courses' in Demand Forecasting. *European Journal of Operational Research*, 237(1):152 – 163.

Sapankevych, N. and Sankar, R. (2009). Time series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, 4(2):24–38.

Schwarz, G. E. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, (6):461–464.

Syntetos, A., Babai, Z., Boylan, J., Kolassa, S., and Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, 252(1):1–26.

Syntetos, A. and Boylan, J. (2001). On the bias of intermittent demand estimates. *International Journal of Production Economics*, 71:457–466.

Syntetos, A., Boylan, J., and Disney, S. (2009). Forecasting for inventory planning: A 50-year review. *Journal of the Operational Research Society*, 60:149–160.

Taylor, C. J., Pedregal, D. J., Young, P. C., and Tych, W. (2007). Environmental Time Series Analysis and Forecasting with the Captain Toolbox. *Environmental Modelling & Software*, 22(6):797–814.

Trapero, J., Kourentzes, N., and Fildes, R. (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society*, 66:299–307.

Trapero, J. and Pedregal, D. (2016). A novel time-varying bullwhip effect metric. an application to promotional sales. *International Journal of Production Economics*, In press.

Trapero, J., Pedregal, D., Fildes, R., and Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting*, 29(2):234–243.

Villegas, M. A. and Pedregal, D. J. (2016). Sspace: A toolbox for state space modelling. *Working paper, UCLM, submitted to the Journal of Statistical Software*.

Winters, P. R. (1960). Forecasting Sales by Weighted Moving Averages. *Management Science*, (6):324–342.

Young, P. C. (2011). *Recursive Estimation and Time-Series Analysis*. Springer-Verlag, 2nd edition.

Young, P. C., Pedregal, D. J., and Tych, W. (1999). Dynamic Harmonic Regression. *Journal of Forecasting*, 18(6):369–394.